Sentiment Analysis Report

Author: Caleb Nicholas Youhanna

4.1 Description of the Dataset Used

The dataset used in this project is the Datafiniti **Amazon Consumer Reviews of Amazon Products (May 2019)** dataset. It contains customer reviews of Amazon products along with associated metadata.

For the purpose of sentiment analysis, only the reviews.text column was used. This column contains written customer feedback and serves as the feature variable for analysing sentiment. All records with missing review text were removed prior to analysis to ensure data quality.

4.2 Details of the Preprocessing Steps

Several preprocessing steps were applied to prepare the text data for sentiment analysis:

- Rows with missing values in the reviews.text column were removed.
- All review text was converted to lowercase.
- Leading and trailing whitespace was removed.
- Stop words and punctuation were removed using spaCy's built-in linguistic features.
- The cleaned text was processed using the en_core_web_md spaCy model with the TextBlob extension enabled.

These steps ensured that the text data was standardised and suitable for sentiment analysis.

4.3 Evaluation of Results

The sentiment analysis model successfully identified sentiment polarity for customer reviews. Reviews with positive language produced positive polarity scores, while negative language resulted in negative polarity scores. Neutral reviews produced polarity scores close to zero.

Testing the model on sample reviews demonstrated that it was generally effective at classifying sentiment in a way that aligns with human interpretation. Polarity scores also provided an indication of the strength of sentiment expressed in each review.

4.4 Insights into the Model's Strengths and Limitations

Strengths

- Simple and interpretable sentiment scores using polarity values.
- Efficient processing of large volumes of text using spaCy.
- Ability to compare similarity between reviews using word embeddings.

Limitations

- The model may struggle to detect sarcasm or context-dependent sentiment.
- Neutral sentiment can be difficult to interpret accurately.
- Results depend on pretrained models and may not fully capture domain-specific language.